Genomics Data 11 (2017) 9–16

Contents lists available at ScienceDirect

Genomics Data

journal homepage: www.elsevier.com/locate/gdata

Traditional *versus* 3′ RNA-seq in a non-model species

Sophie Tandonnet [a], Tatiana Teixeira Torres [a,b,*]

[a] Department of Genetics and Evolutionary Biology, Institute of Biosciences, University of São Paulo, São Paulo, SP, Brazil
[b] Research Center on Biodiversity and Computing (BioComp-USP), Universidade de São Paulo (USP), São Paulo, SP, Brazil

ARTICLE INFO

Article history:
Received 23 August 2016
Received in revised form 1 November 2016
Accepted 2 November 2016
Available online 18 November 2016

Keywords:
Gene expression profiling
Length bias
Transcriptome
Cochliomyia hominivorax
Calliphoridae

ABSTRACT

One limitation of the widely used RNA-seq method is that long transcripts are represented by more reads than shorter transcripts, resulting in a biased estimation of expression levels. The 3′ RNA-seq method, which yields only one sequence per transcript, bypasses this limitation. Here, RNA was extracted from two samples, in which we expected to find differentially expressed genes. Each was processed by both traditional and 3′ RNA-seq protocols. Both methods yielded similar differentially expressed genes and estimated expression levels in a comparable way, confirming they both represent valid tools for RNA-seq analysis. Notably, however, we identified more differentially expressed transcripts with the 3′ RNA-seq method, suggesting a greater power to detect expression variation using this method. Hence, when little genomic information is available for the species studied, the standard RNA-seq presents a better cost-benefit compromise, whereas for model species, the 3′ RNA-seq method might more accurately detect differential expression.

1. Introduction

The RNA-seq method is a powerful tool allowing for functional genomic studies at the transcriptional level. It consists of the deep sequencing of the RNA (total or fractionated) of an individual or tissue at a certain time and condition. This approach enables the comparative assessment of the level of expression for each gene between different samples. By comparing the RNA expression profiles among samples, it is possible to identify differentially expressed (DE) genes that might explain the phenotypic differences observed between the samples.

In this study, we compared two RNA-seq methods: the standard RNA-seq and the 3′ RNA-seq that is expected to give more accurate levels of expression by solving some of the biases inherent in the classic RNA-seq method. With the standard RNA-seq method, the extracted mRNA is randomly sheared and the fragments are converted into a cDNA library. The cDNA fragments are then sequenced by one of the next-generation sequencing technologies. The total number of reads (cDNA fragments sequenced) corresponding to a given transcript is proportional to the level of expression of the corresponding gene [11]. However, one of the limitations of the standard RNA-seq strategy lies in the fact that longer transcripts are broken into more fragments than are shorter ones. This creates a statistical bias, as longer transcripts will be represented by more reads than those produced by the shorter transcripts. Consequently, the detection of DE is more likely to be over-represented for long transcripts and under-represented for shorter ones, which are at a statistical disadvantage [11]. To minimize this bias, the levels of expression (number of reads corresponding to a certain transcript) can be corrected by the size of the transcript. However, in the case of non-model species, this information is most likely to be unavailable. The correction can then be done by using the contig size from the *de novo* reconstruction of the transcript (based on the reads) or by employing the transcript sizes of a closely related model species. Nevertheless, this correction does not entirely solve the problem owing to the transcript size, as the sampling is higher for longer transcripts [11].

The 3′ RNA-seq method [15] was conceived to bypass these limitations. This method consists of sequencing only one fragment per transcript in the 3′ region. By using this strategy, regardless of the transcript length, the levels of expression can be estimated directly by the number of reads corresponding to a certain transcript, as a single fragment per mRNA molecule is sampled (Fig. 1).

In this paper, we compare both RNA-seq methods at the different steps of an RNA-seq analysis to clarify their advantages, disadvantages, and complementarities for a non-model species, *Cochliomyia hominivorax*, the New World screw-worm fly. This species is one of the most important myiasis-causing fly of the neotropical region and is responsible for severe economic losses. During the last decades, *C. hominivorax* populations were mainly controlled by applying

* Corresponding author at: Department of Genetics and Evolutionary Biology, IB-USP, Rua do Matão, 277, 05508-090 São Paulo, SP, Brazil.
E-mail address: tttorres@ib.usp.br (T.T. Torres).

http://dx.doi.org/10.1016/j.gdata.2016.11.002

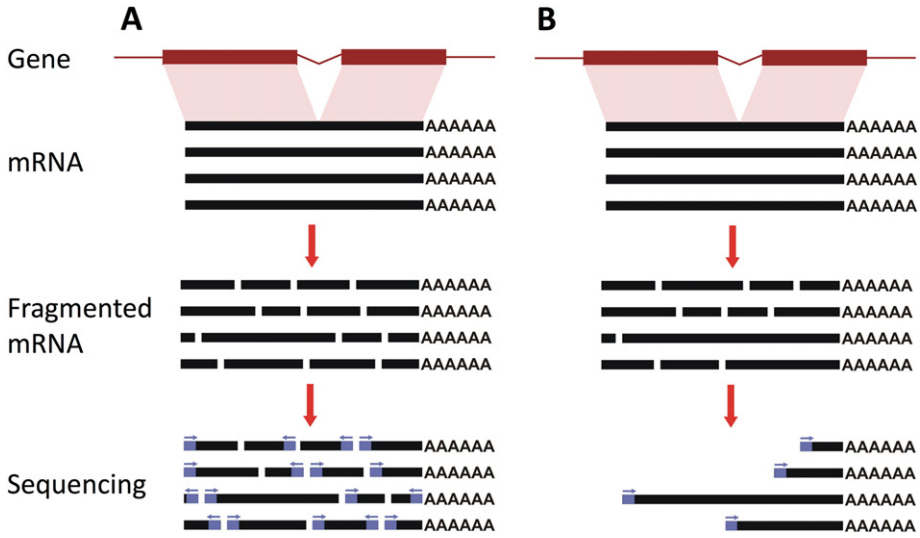

Fig. 1. Overview of the methods used to generate the RNA-seq libraries. (A) In the classic RNA-seq procedure, the RNA is fragmented and converted into cDNA using small primers of random sequence. (B) In the 3′ RNA-seq library, the mRNA molecules are randomly fragmented, generating fragments of different lengths. After fragmentation, only the 3′ portion of an mRNA molecule is selected using poly-T oligonucleotide baits attached to magnetic beads. The selected fragments (one per molecule) are then directionally sequenced.

organophosphate (OP) insecticides, but because of this constant selective pressure, resistant lineages have been strongly selected, complicating the management of this species [3,4]. In this context, the RNA-seq methods were used to discover the genes possibly involved in OP resistance.

2. Materials and methods

2.1. C. hominivorax populations

We used a laboratory colony of *C. hominivorax* composed of susceptible and known OP resistant individuals (Gly137Asp and/or Trp251Ser mutations in the esterase E3 gene), collected in Caiapônia, GO, Brazil. The colony was maintained according to standard protocols [2]. For the resistant condition, a sample from the laboratory population was treated with the OP insecticide dimethyl 2,2-dichlorovinyl phosphate; $C_4H_7Cl_2O_4P$ (dichlorvos) at 20 mg/l, a concentration lethal for 90% of the population (LC90). The insecticide was directly mixed into the medium consisting of fresh ground beef supplemented with blood and water (2:1). A total of 500 L2 instar larvae were fed on the insecticide-containing medium for 24 h. The surviving individuals (Resistant sample) were collected for the RNA extractions. The individuals of the control condition were simply sampled from this laboratory population and fed on the medium without the insecticide.

2.2. RNA extraction

RNA extractions followed previously utilized procedures [2]. Total RNA of resistant and control *C. hominivorax* larvae were extracted separately using TRIzol (Invitrogen) from the whole bodies of 87 larvae, 42 from the resistant and treated group and 45 from the control group. DNase I (Invitrogen) was used to remove genomic DNA contamination and the mRNA-enriched samples were further purified using Nucleospin RNA Clean-up columns (Macherey Nagel). RNA quantification was performed using the Qubit Quantitation Platform fluorometer (Invitrogen).

2.3. RNA-seq experiments

The extracted RNA was processed separately according to the two RNA-seq protocols. In the classic RNA-seq procedure, the RNA fragments resulting from the random breakage of the transcripts were converted into a cDNA library using the mRNA-Seq Sample Prep Kit (Illumina). Small primers (6 nt) of random sequence were used to produce the cDNA fragments. Specific adapter sequences (ACGTT and TGCAT for the control and resistant conditions, respectively) were prefixed to the cDNA fragments. These barcoded control and resistant cDNA sequences were then pooled prior to sequencing. Library preparation was performed independently twice on the same samples (technical replicates).

The 3′ RNA-seq library was constructed by Fasteris (Switzerland) using the procedure adapted from a previous study [15]. In this method, 4 μg total extracted RNA for each sample (control and resistant) was used to create the 3′ RNA libraries. A 3′ RNA library contains only those RNA fragments possessing a polyA tail. For its construction, the mRNA- Seq Sample Prep Kit (Illumina) was modified to select the 3′ RNA fragments. Briefly, the mRNA molecules were fragmented at a high temperature (80 °C) by divalent cations using the fragmentation buffer. The polyA mRNA fragments were purified using poly-T oligonucleotide baits attached to magnetic beads. After the selection of polyA fragments, the mRNA-seq Sample prep Kit (Illumina) protocol was followed according to the manufacturer's instructions. Consequently, we obtained one polyA-fragment per transcript molecule, which allowed us to directly estimate the expression level of the transcripts. Resistant and control samples were pooled prior to sequencing by Fasteris using the Illumina HiSeq100 system (single-reads of 100 bp).

2.4. Preprocessing of the reads

For a thorough comparison between the RNA-seq methods, we sampled the same number of raw reads obtained by both methods for each condition. Since the sequencing based on the cDNA obtained by the standard RNA-seq method yielded fewer reads (15,427,065 for control, 17,021,595 for resistant), we sampled those numbers of raw reads from both 3′ RNA-seq read populations (35,574,183 for control, 46,322,457 for resistant). The sampling was performed using the function “FastqSampler” from the R package “ShortRead” [10].

To eliminate poor quality regions of the sequences, we used the program fastq_quality_trimmer from the fastx toolkit suite (http://hannonlab.cshl.edu/fastx_toolkit/). We used the default quality score threshold of 20 and removed the sequences shorter than 20 bases after the trimming had being completed.

The quality of the reads was assessed before and after the trimming step using the program FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). This program allowed us to visualize our data and was used to confirm the good quality of the reads prior to usage.

2.5. De novo assembly of the C. hominivorax reference transcriptome

De novo transcriptome assemblies were performed using the program Trinity [6], wherein we set the minimum contig length to 50 bp. Three independent assemblies were performed. The first two assemblies were implemented, separately, on the collapsed reads obtained by the two RNA-seq methods in order to compare the performance of both methods in recovering full-length transcripts. Collapsing consisted in finding identical reads in the "trimmed" sequence files and retaining only one copy by using an in-house Perl script. Using the collapsed processed reads reduced markedly the processing time without any loss of accuracy during contig assembly.

For the third assembly, we used all available reads from the Illumina sequencing (standard and 3′ RNA-seq methods) as well as from published 454-pyrosequencing data [2]. This was done to generate a more complete database of *C. hominivorax* transcripts.

The assemblies generated for both methods were evaluated for their accuracy and completeness. The RSEM-EVAL component of the DETONATE package [9] was used to measure the accuracy of the assemblies. RSEM-EVAL evaluates the compactness of an assembly and the support of the assembly from the RNA-Seq data. The REF-EVAL contig- and nucleotide-level measures were compared between the assemblies using the 3′ RNA and the traditional method. The Core Eukaryotic Genes Mapping Approach (CEGMA) (version 2.5) [12] was used to evaluate the completeness of the assemblies.

2.6. Annotation

We performed a local blastn using the program BLAST + [1] to annotate the assembled contigs using a reference database containing *Drosophila melanogaster* transcripts, coding DNA sequence (CDS), genes, and extended genes (1000 nucleotides up-and downstream of the gene) from Flybase (http://flybase.org/, Release 5.50). The BLAST output was then parsed using an in-house Perl script that allowed us to set the minimum percentage of identity with the database sequence at 50% and the minimum percentage of query sequence involved in the alignment at 50%. The E-value was set at 1e-4.

2.7. Counting of the poly(A)-rich reads

The number of the processed reads containing a minimal percentage of As (50 to 100%) was counted using an in-house Python script. Additionally, our script allowed us to retain the reads (in fastq format) that satisfied the minimal percentage of As set by the user. We then attempted to align these reads against a set of *C. hominivorax* contigs assembled using all available reads using the program Bowtie [7]. The Bowtie results were then parsed and the ambiguous reads (*i.e.*, reads aligning against more than one contig) were discarded.

2.8. Read alignment

Processed reads originating from both RNA-seq methods were mapped to the correspondent set of *C. hominivorax* contigs. The read alignments were performed using the program Bowtie [7] for both methods and conditions with the default settings.

To compare the distribution and coverage of the reads from the different methods along genomic sequences, we also aligned the processed reads against the mitochondrial genome of *C. hominivorax* [8] and plotted the results as previously described [14].

2.9. Abundance estimation of the transcripts for each condition

Based on the output of Bowtie, we counted the reads that aligned against each contig *via* an in-house Perl script for each condition and for each RNA-seq method. The Bowtie parser discarded the ambiguous reads that aligned to more than one contig with the same score (same number of mismatches). When a read aligned against more than one contig isoform or "allele contig" (Trinity outputs such variants) the read was counted as part of the common region of the transcript isoform. This counting strategy allowed a finer comparison between conditions at the level of the transcript variant. All scripts used for data analysis are available upon request.

2.10. Differentially expressed transcripts and annotation

The R package EdgeR [13] of the Bioconductor repository was used to identify the DE contigs between the conditions for both methods. The contigs that did not display at least one (read) count per million in both the resistant and control libraries were excluded from the analysis (filtering option).

Since biological replicates were not available, we estimated the biological variation to be 0.2. Our estimation was based on examples of biological variation from the EdgeR User's Guide and the source of the *C. hominivorax* sample. The fly colony was maintained in the laboratory for nearly one year prior to the experiments and it is not an outbred population nor yet a truly inbred colony. Hence, the adopted variation (0.2) was between the default variation for outbred populations (0.4) and the default variation for inbred populations (0.1).

As many of the DE contigs were not annotated by using the *Drosophila* sequences, we re-annotated them *via* remote blast against the GenBank database. We used the default "nr" (nucleotide, non-redundant) database and the tblastx algorithm to maximize the chances of a significant alignment. The "nr" database regroups all the sequences from GenBank, European Molecular Biology Laboratory (EMBL), DNA Data Bank of Japan (DDBJ), and Protein Data Bank (PDB) sequences. In addition, we only retained the blast alignments exhibiting a minimum of 50% identity with the database sequence and with a minimum of 20% of the query sequence involved in the alignment. The E-value was set at 1e-4.

To better compare the top 10 up- and down-regulated contigs of both methods, we attempted to identify the top DE contigs not annotated by remote blast. To do so, we used the program Hmmer through the web server [5], conserving the default settings along with the UniProtKB protein database (UniProtKB/Swiss-Prot and UniProtKB/TrEMBL).

Table 1
Number of raw, processed, and collapsed reads for each RNA-seq method and sample condition.

	Standard RNA-seq		3′RNA-seq	
	Control	Resistant	Control	Resistant
Raw reads	15,427,065	17,021,595	15,427,065	17,021,595
Processed reads	14,545,819	16,103,624	15,389,704	16,973,215
(% of reads discarded during trimming and clipping)	(5% discarded)	(5% discarded)	(0% discarded)	(0% discarded)
Number of poly (A) sequences from the processed sequence files	81	36	563,870	875,928
Collapsed reads	6,987,224	9,037,903	5,018,226	5,911,574
(% of the processed reads discarded)	(52%)	(44%)	(68%)	(66%)

Poly(A) sequences are those consisting exclusively of As.

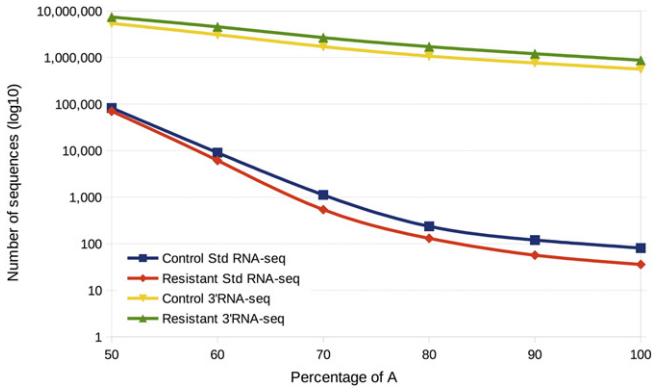

Fig. 2. Number of processed reads with an increasing proportion (percentage) of As. The 3′ RNA-seq method yields many more reads primarily composed of As than does standard (Std) RNA-seq.

To test if the average length of the DE contigs differed significantly between each method, we performed a non-parametric Wilcoxon rank sum test with continuity correction using the function "wilcox.test" of the statistical platform R.

2.11. Quantitative PCR (qPCR)

We used the qPCR data from the study of Carvalho et al. [2] as the conditions for the bioassays were the same as in our study. qPCRs were performed on 18 candidate genes (acetylcholinesterase, alpha-esterases 7, 8, and 9, serineprotease 7, glutathione S transferases D1, E5, and S1, and *Cyp4ac1*, *Cyp4c3*, *Cyp4d2*, *Cyp6a14*, *Cyp6a9*, *Cyp6d4*, *Cyp6g1*, *Cyp6v1*, *Cyp9f2*, and *Cyp12a4*) and we recovered the data of all these genes for our analysis.

2.12. Correlations and comparisons between determined expression levels

Correlations between the levels of expression of the contigs generated for each RNA-seq method were calculated using Spearman's rank correlation rho (function "cor.test" in R). This was done separately for each condition, control and resistant. The counts were corrected by the length of the contigs for the standard RNA-seq and by the size of the library for both methods. To avoid taking into account the inaccuracy associated with low count numbers, we only considered the contigs with a minimum of ten reads in at least one RNA-seq method.

For the comparison between RNA-seq and qPCR methods, we corrected the mRNA counts by the size of the library (number of processed reads) and calculated the Spearman's rank correlation between the log2 (control expression/resistant expression) for the RNA-seq methods and the ddCT for qPCR.

3. Results

3.1. RNA-seq data and preprocessing

After sampling the same quantity of raw reads for each RNA-seq method (Fig. 1), we trimmed the regions presenting a low per-base quality (<20) and discarded the resulting processed reads shorter than 20 bp. With these parameters, we discarded more standard RNA-seq than 3′ RNA-seq reads (Table 1). However, the inverse pattern occurred when we collapsed the files of processed reads; *i.e.*, when we removed all the identical reads, retaining only one copy (Table 1). Approximately 67% of the processed 3′ RNA-seq reads were discarded during the collapsing process whereas only about 48% of the standard RNA-seq reads were removed. This was partly due to the higher number of poly(A) sequences obtained by the 3′ RNA-seq method (Fig. 2), as poly(A) sequences of the same length were collapsed into a single read.

Table 2
Summary statistics of assembly results using the program Trinity.

	RNA-seq	3′RNA-seq
Number of contigs longer than 50 bp	76,089	90,752
First quartile of contig lengths	90 bp	98 bp
Median contig length	119.0 bp	143.0 bp
Average contig length (SD)	280.2 bp (559.52)	181.8 bp (139.87)
Third quartile of contig lengths	218.0 bp	226.0 bp
Longest contig (bp)	24,277	5866

SD, Standard deviation.

3.2. Assembly and annotation of the transcriptome

The *de novo* assembly of the reads obtained by each RNA-seq method, aiming at reconstructing the original transcripts, was more efficient using the standard RNA-seq reads than those from 3′ RNA-seq (Table 2). The major difference between the methods was observed for the length of the contigs assembled. Using the standard RNA-seq reads, the assembly yielded longer contigs (up to >20,000 bp) whereas with the 3′ RNA-seq reads, the length of the assembled contigs quickly dropped, attaining a maximum of approximately 5000 bp (Fig. 3). This was an expected result as the 3′ RNA-seq method only retains the 3′ region of the transcripts while the standard RNA-seq method produces reads distributed along the transcript length. It is noteworthy that, even in the 3′ RNA-seq method, long transcripts are obtained owing to the random fragmentation of fragments; each mRNA molecule will be fragmented at a different site and the 3′ fragments selected will have different lengths (Fig. 1).

The transcript assemblies were also compared in terms of their accuracy. We computed the RSEM-EVAL scores for the transcriptome assemblies generated for each method and the results were very similar, −638,724,079.59 for the assembly performed using reads from the traditional method and −664,840,190.97 with reads from the 3′ RNA-seq method. We also compared their completeness using the CEGMA pipeline. The traditional method recovered 83.47% of the 248 ultra-conserved core eukaryotic genes (CEGs) whereas 3′ RNA-seq recovered only 27.02%. Notably, the 3′ RNA-seq method recovered the 3′ region of transcripts rich in untranslated regions that are more divergent and, therefore, are more difficult to be identified.

This difference among the methods was confirmed by aligning the reads originated by both methods against the available mitochondrion genome of *C. hominivorax* (Fig. 4). This analysis also highlighted another important feature of the 3′ RNA-seq method: that it results in directional reads; *i.e.* the orientation of the read is known.

Annotation of the assembled contigs was accomplished by local blasts (blastn) performed against a *D. melanogaster* database. The contigs assembled from the standard RNA-seq reads were more efficiently annotated (by approximately 12%) than those from the 3′

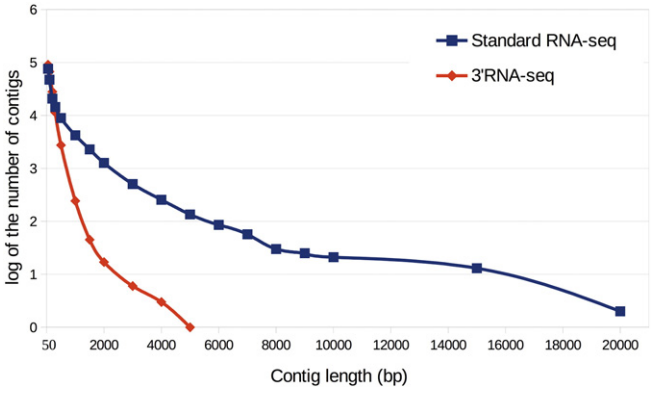

Fig. 3. Evolution of the number of contigs (log10) with respect to their length (bp) for both RNA-seq methods.

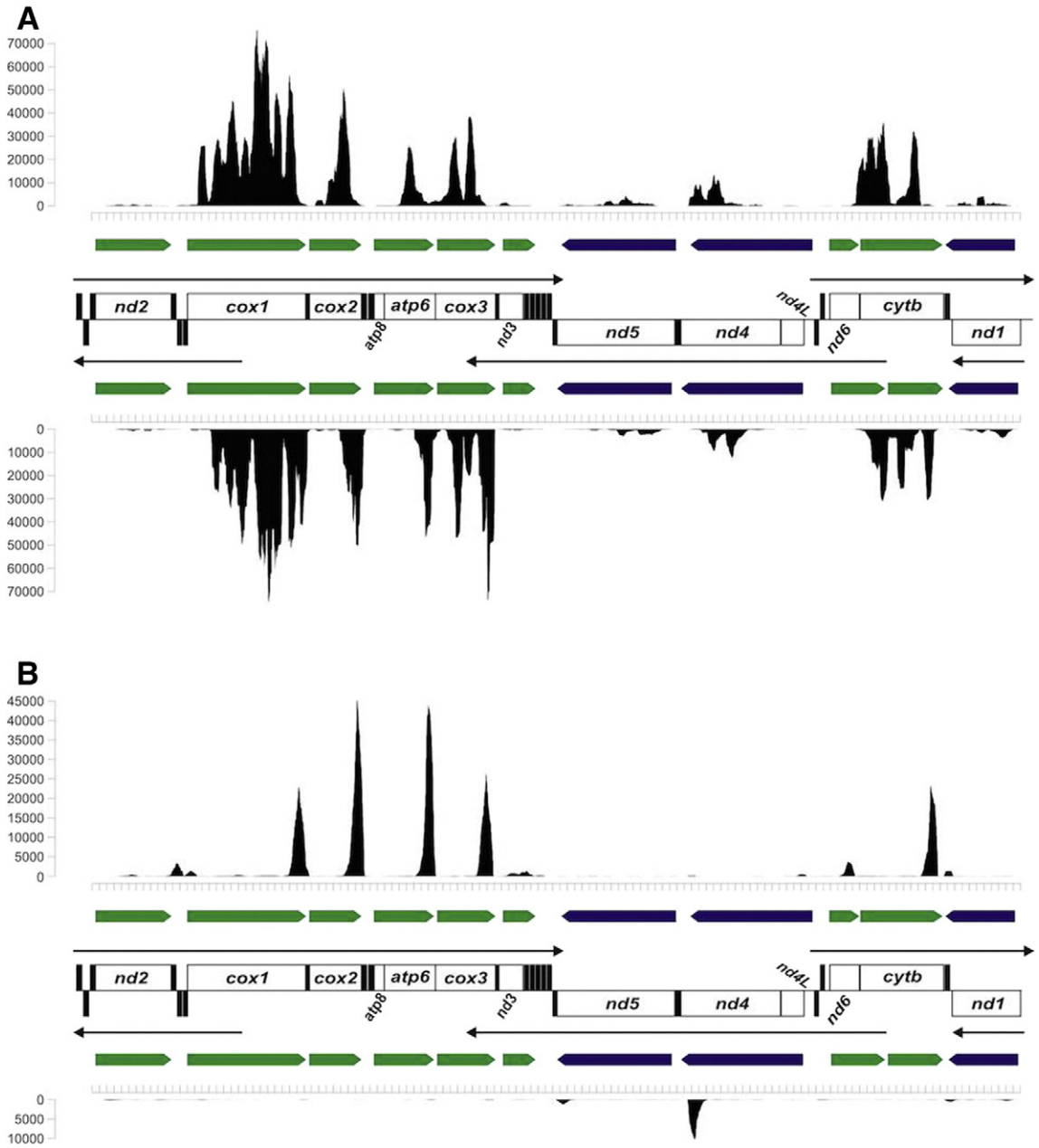

Fig. 4. Alignment of the reads obtained by standard RNA-seq (A) and 3′ RNA-seq (B) against the mitochondrial genome of *C. hominivorax*. The histograms in black represent the coverage of each base of the mitochondrial genome. In both (A) and (B), the upper histograms represent the reads aligned against the "+" strands and the lower ones represent the alignments that occurred with the "−" strand. The genes coding proteins are represented by the unfilled boxes and the transporter RNAs by small black-filled boxes. The coding units are represented by the black arrows and the mature transcripts by filled arrows, with those of the "+" strand in green and those of the "−" strand in blue. The ribosomal rRNA genes were not included as their expression level is much higher than that of the other transcripts. Only the control samples are represented since the alignments performed with the resistant samples gave very similar results.

RNA-seq method (Table 3). The proportion of shared annotation between both methods was around 59% of the total annotations of the standard RNA-seq contigs and 78% of the 3′ RNA-seq contigs.

The large quantity of reads mainly composed of As found in the 3′ RNA-seq output might in practice cause a substantial loss of sequence data. To assess this possibility, we aligned the «A-rich» reads against the set of *C. hominivorax* contigs assembled using all available data. Many reads (over 70% of the 3′ RNA-seq processed reads composed of at least 50% As) did not align against the contigs and therefore did not contribute to the expression level of their corresponding gene.

Notably, we observed that the percentage of aligned reads did not decrease as had been expected as the percentage of As in the reads increased (Fig. 5). The percentage of aligned and unambiguous reads with a high proportion of As stabilized and even slightly increased under certain conditions. This suggests that some assembled contigs were mostly composed of As and therefore were devoid of relevant information owing to the impossibility to annotate them.

Table 3
Number of annotated contigs.

	Standard RNA-seq	3′RNA-seq
Total number of contigs	76,089	90,752
Number of contigs with no hit	2395	8032
Number of discarded hits	54,127	70,154
Number of sequences aligned	19,567 (25.72%)	12,566 (13.85%)
Number of sequences not aligned	56,522	78,186

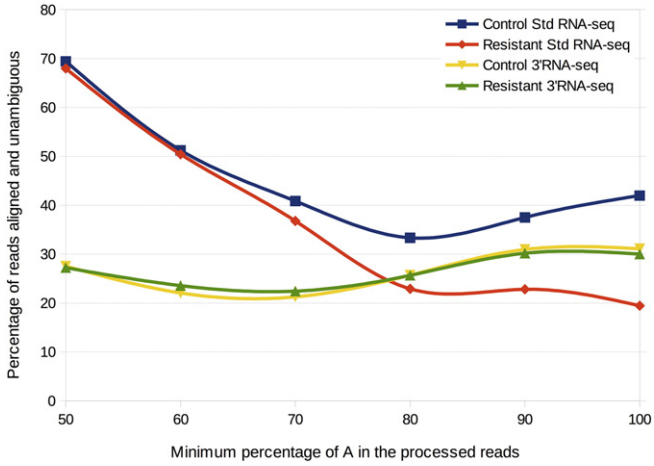

Fig. 5. Percentage of reads with an increasing proportion of As aligned against a set of *C. hominivorax* contigs. A read is said to be aligned uniquely if it has aligned only once or if the second best alignment has a different number of mismatches. As the poly-A lengths are different between the contigs, a read with 100% As can be uniquely mapped. Std, standard RNA-seq.

3.3. Comparison of the expression profiles

We compared both RNA-seq methods based on the performance of the read alignments. To do so, the processed reads from each condition and RNA-seq method were mapped against a common database of *C. hominivorax* contigs assembled with all available transcriptomic data (ours and those from Carvalho et al. [2]. The results showed that a larger number of 3′ RNA-seq reads failed to align (~47%) when compared to the standard RNA-seq reads (~24% alignment failure). This result was observed for both control and resistant conditions (Table 4) and might be due to the large amount of poly(A) sequences present in the 3′ RNA-seq processed reads. For both methods, almost all aligned reads were unambiguous (Table 4).

To assess the significant differences in expression between conditions for each contig, we used the EdgeR package [13]. EdgeR applies an exact test to check for DE between conditions by fitting a negative binomial model for each transcript. Using this method, we could identify the probable DE contigs between the control and resistant conditions for both RNA-seq methods. We found more DE contigs with 3′ RNA-seq (~1.57 times more) than with standard RNA-seq (Table 5 & Fig. 6), suggesting that we potentially had greater power to detect DE contigs with the 3′ RNA-seq method.

As many of the DE contigs were not annotated by using the *Drosophila* database, a second effort was made to identify these sequences by remote blast (tblastx) against the GenBank database. The results obtained were compared in numeric and biological terms. A greater number of DE contigs derived from the standard RNA-seq protocol could be annotated (~53%) when compared to those from 3′ RNA-seq (~36%). As 3′ RNA-seq favors the 3′ regions that are enriched in 3′UTRs, that are more difficult to annotate. The majority (75–79%) of the successfully annotated DE contigs shared the same identifier (best hit) between both methods (Fig. 7).

Table 5
Summary of the number of DE contigs in the Resistant condition when compared to the Control condition.

	Standard RNA-seq	3′ RNA-seq
Number of contigs similarly expressed	12,026	14,180
Number of contigs over-expressed	104	172
Number of contigs under-expressed	157	239
Number of differentially expressed (DE) contigs	261	411

Exact test, false discovery rate (FDR) < 0.05.

We also compared the top 10 most significant up- and down-regulated contigs of both methods (Supplementary Tables S1 to S4). The cytochromes P450 cyp6g4 and cyp6a28, the collapsin response mediator protein (CRMP) and a glucuronosyl transferase were among the top 10 down-regulated contigs in both RNA-seq methods (Supplementary Tables S1 and S2). Several genes involved in metabolic processes (a Glucose/ribitol dehydrogenase, an alcohol dehydrogenase, and a hydrolase) only appeared in the standard RNA-seq top 10 down-regulated contigs (Supplementary Table S1). In the top 10 up-regulated contig lists (Supplementary Tables S3 and S4), an odorant binding protein gene and a chitin binding peritrophin-A gene as well as two unknown genes were identified in both methods. In the top 10 up-regulated contigs, we also found two different genes involved in immunity, a sarcotoxin in the standard RNA-seq listing and a defensin-A in the 3′ RNA-seq group. Although many up-regulated contigs among the standard RNA-seq top 10 corresponded to heat shock proteins (*HSP70*), none such were found in the 3′ RNA-seq top 10. However, various contigs of the 3′ RNA-seq top 10 up- or down-regulated lists were unknown or corresponded to uncharacterized proteins, which limited our qualitative comparison.

3.4. Correlations between contig expression levels

The Spearman's rank correlation coefficient between the levels of expression of the contigs assessed for each RNA-seq method was calculated separately for each condition. The Spearman's rank correlation rho between the methods was 0.53 for the control condition and 0.50 for the resistant condition. This correlation was done without assessing the identity of the contigs. Hence, different contigs from the same gene would be considered completely different entries. It would be possible to circumvent this problem by aligning the reads against an annotated reference genome; however, the complete genome sequence of *C. hominivorax* is not currently available. On the other hand, the sequence and annotation of the mitochondrial genome of *C. hominivorax* is available and we therefore also assessed the correlation between methods for the mitochondrial genes as previously described. We obtained correlations of 0.84 and 0.83 (Spearman's rank correlation rho) for the control and the resistant conditions, respectively.

For 18 genes (Table S5), we verified whether the levels of expression obtained through the RNA-seq methods were consistent with those obtained by quantitative PCR (qPCR). Inconsistencies were observed for the *Cyp6a14*, glutathione S transferase S1, and alpha-esterase7 genes

Table 4
Read alignment and parsing data in each condition (Control and Resistant) and for both RNA-seq methods.

		Standard RNA-seq		3′ RNA-seq	
		Control	Resistant	Control	Resistant
Bowtie alignment	Trimmed reads (non-poly(A))	14,545,819 (14,545,738)	16,103,624 (16,103,588)	15,389,704 (14,825,834)	16,973,215 (16,097,287)
	Reads that failed to align	3,102,159 (21.33%)	4,321,671 (26.84%)	6,899,084 (44.83%)	8,311,612 (48.97%)
	Reads with at least one reported alignment	11,443,660 (78.67%)	11,781,953 (73.16%)	8,490,620 (55.17%)	8,661,603 (51.03%)
Bowtie parser	Unambiguous aligned reads	11,330,380 (~99%)	11,676,453 (~99%)	8,375,287 (~98.6%)	8,485,019 (~98%)
	Discarded aligned reads (ambiguous)	113,280	105,500	115,333	176,584
	Number of contigs	59,728	64,644	56,415	63,211

Number of alignments and reads processed by the program Bowtie and number of contigs and unambiguous/ambiguous reads.

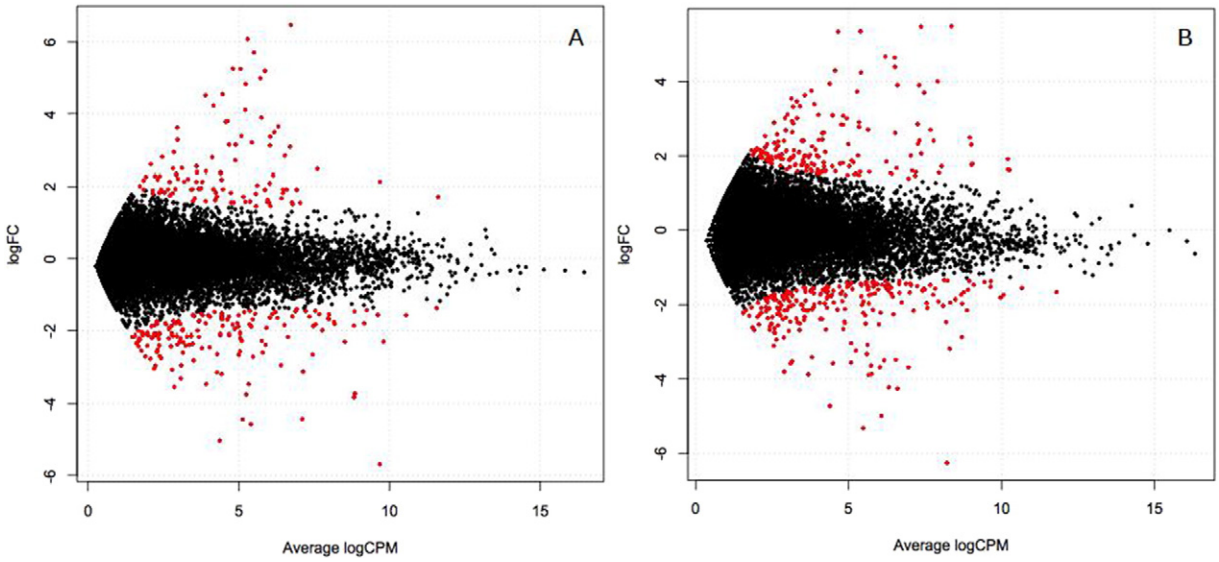

Fig. 6. Fold Change (FC) plots showing the contigs differentially expressed (DE, red dots) in the resistant sample when compared to the control sample (Exact test, adjusted p-value < 0.05). (A) FC plot obtained from the classical RNA-seq data and (B) from the 3′ RNA-seq data. CPM (counts per million) represent how many reads per million mapped unambiguously against a contig; FC is the difference between the abundance of reads (*i.e.*, the level of expression) in the resistant and control samples. Up-regulated DE contigs correspond to those with a positive logFC while the down-regulated contigs have a negative logFC.

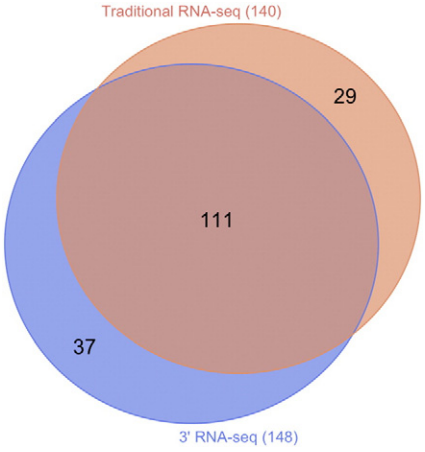

Fig. 7. Proportion of annotated DE contigs found in common or not between both RNA-seq methods. For both methods, annotation was performed by remote blast using the non-redundant database of NCBI. The contigs sharing the same annotation (same best hit number) in both methods were counted in common ("VennDiagram" package of the R platform).

between the RNA-seq and the qPCR methods (Fig. 8). Overall, the qPCR expression ratios were closer to those obtained using standard RNA-seq (Spearman's rank correlation rho 0.8333333) than to the 3′ RNA-seq ratios (rho 0.547619). However, low read count for the analyzed genes were mostly observed in the 3′ RNA-seq method, which might be explained by the fact that the 3′ RNA-seq reads corresponded mostly to 3′UTR regions that were absent in the regions of sequences that had been used to generate the PCR primers.

4. Discussion

The 3′ RNA-seq method was first used in 2008 incorporating the 454 sequencing technology as a new approach to measure gene expression by performing massively parallel sequencing [15]. Considering the current boom of RNA-seq and realizing the biases associated with this methodology, we adapted the 2008 method for sequences generated using the Illumina platform. We then performed a comparison between the outputs of standard and 3′ RNA-seq to understand some of the differences, similarities, and complementarities between both methods.

We identified similar DE genes from both RNA-seq techniques, which confirmed that both methods represent valid tools to investigate

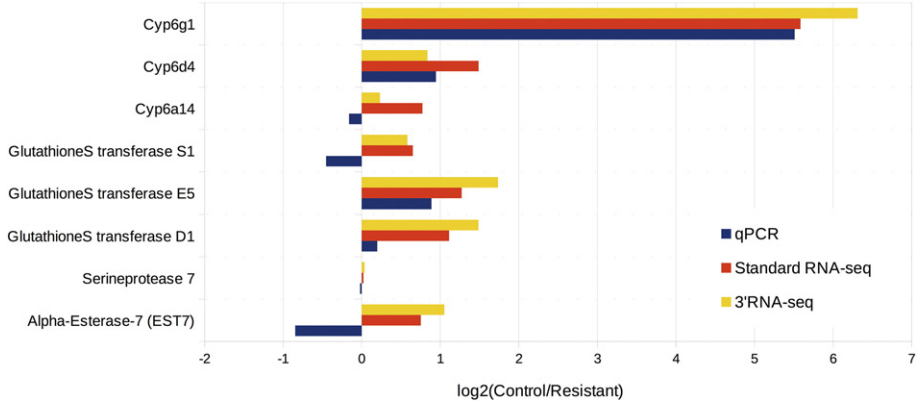

Fig. 8. Comparison of the qPCR, standard-RNA-seq, and 3′ RNA-seq expression ratios. The levels of expression were corrected by the size of the library for both RNA-seq methods. For the RNA-seq methods, only the genes displaying a minimum raw read count of 10 in at least one condition were considered.

DE. It is worthy to note, however, that we obtained a substantially larger number of DE contigs using 3′ RNA-seq. This suggests that 3′ RNA-seq might have more power to detect DE. Nevertheless, few of the identified DE contigs could be annotated. In particular, many of the 3′ RNA-seq reads correspond 3′UTRs, which are prone to accumulate sequence changes. Being less conserved, 3′UTR are harder to align and annotate. For a model species with an available and well annotated genome, we therefore would expect a higher rate of annotation and therefore, a better performance of the 3′ RNA-seq method. As expected, we found that with standard-RNA-seq, a better *de novo* transcriptome assembly could be completed. The standard RNA-seq reads were distributed along the transcripts whereas only the 3′ regions of the transcripts were sequenced with the 3′ RNA-seq method. Consequently, most of the transcript is lost with 3′ RNA-seq, making *de novo* transcriptome assembly difficult.

The relatively low values of correlation (~0.50) between the expression levels determined by each method might be due to the bias created during the transcriptome assembly. The *C. hominivorax* contig set we used was mostly constructed with the standard RNA-seq reads since the 3′ RNA-seq reads were restricted to the 3'region of the transcript (Fig. 5). Supporting this hypothesis, we found high correlations (~0.80) between the expression levels of both methods for the mitochondrial genes, which were available for our species. These high correlations confirm that both methods assess expression levels in a similar fashion.

Although we predicted that the 3′ RNA-seq technique would yield more accurate levels of expression, we were unable to confirm this conjecture through our study. *De facto*, the qPCR expression levels were closer to those from the standard RNA-seq method than from 3′ RNA-seq; however, only a few genes had been assayed by qPCR. The standard RNA-seq method is known to present a statistical bias owing to the tendency of longer fragments to be broken into a larger number of fragments [11]. We observed this when comparing the average length of the DE contigs, as those originating from standard RNA-seq were typically longer than the ones derived from 3′ RNA-seq. This observation shows the means by which, with standard RNA-seq, long DE contigs are over-represented in comparison to shorter ones. On the other hand, 3'RNA-seq, which was designed to by-pass this limitation, presented other limitations. Many 3′ RNA-seq reads corresponded to poly(A)-rich sequences that were empty of relevant information. This is likely to reduce assay accuracy since the majority cannot be aligned to their respective contig. A further comparison between both methods performed on model species and/or on a number of genes of varying length might shed light on the overall expression accuracy of the 3′ RNA-seq method.

Notably, owing to the design of the 3′ RNA-seq method, all reads are polarized in the 5′–3′ sense. Because the strand is known, it is possible to study sense/antisense information. This feature is particularly useful for a number of studies that aim at identifying antisense transcripts, untangling overlapping sense/antisense transcripts, or characterizing the antisense transcriptome. As such, the 3′ RNA-seq might be an avant-garde method for transcriptomic studies benefiting from full-length transcript sequencing (as offered, for example, by Pacific Biosciences). With such sequencing technologies, we might expect full length transcript information and expression level accuracy when using the 3′ RNA-seq method.

Overall, our results suggest that when little genomic/transcriptomic information is available for the species studied, the standard RNA-seq method presents a better cost-benefit compromise. It is cheaper and guarantees full length transcript information. On the other hand, for model species with available genome/transcriptomes, the 3′ RNA-seq method may more accurately detect DE, especially for short transcripts, although many reads would be expected to be lost owing to the large amount of poly(A) reads. Furthermore, 3′ RNA-seq might also become a method of choice for RNA-seq studies using full-length transcript sequencing.

Data accessibility

The dataset supporting the results of this article is available in the SRA repository, under the accession number SRP044071.

Acknowledgements

We are very grateful to Magne Osteras, Loïc Baerlocher, and Laurent Farinelli for library preparation and sequencing. We are also thankful to the Torres lab members for helpful comments and suggestions on earlier versions of this manuscript. This work was funded by the São Paulo Research Foundation, FAPESP (grants 2008/58106-0 and 2012/06819-9 to TTT). ST received a fellowship from FAPESP (2013/00243-0) and TTT received a research fellowship from the Brazilian National Council for Scientific and Technological Development (307502/2011-2).

Appendix A. Supplementary data

Supplementary tables (from S1 to S5) summarizing information on the top 10 DE contigs found in both RNA-seq methods and on the genes used for the qPCR analysis. Supplementary data associated with this article can be found in the online version, at http://dx.doi.org/10.1016/j.gdata.2016.11.002.

References

[1] C. Camacho, G. Coulouris, V. Avagyan, et al., BLAST+: architecture and applications. BMC Bioinformatics 10 (2009) 421.
[2] R.A. Carvalho, A.M. Azeredo-Espin, T.T. Torres, Deep sequencing of New World screw-worm transcripts to discover genes involved in insecticide resistance. BMC Genomics 11 (2010) 695.
[3] R.A. Carvalho, T.T. Torres, M.G. Paniago, A.M.L. Azeredo-Espin, Molecular characterization of esterase E3 gene associated with organophosphorus insecticide resistance in the New World screwworm fly, *Cochliomyia hominivorax*. Med. Vet. Entomol. 23 (2009) 86–91.
[4] R.A. de Carvalho, T.T. Torres, A.M. de Azeredo-Espin, A survey of mutations in the *Cochliomyia hominivorax* (Diptera: Calliphoridae) esterase E3 gene associated with organophosphate resistance and the molecular identification of mutant alleles. Vet. Parasitol. 140 (2006) 344–351.
[5] R.D. Finn, J. Clements, S.R. Eddy, HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39 (2011) W29–W37.
[6] M.G. Grabherr, B.J. Haas, M. Yassour, et al., Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29 (2011) 644–652.
[7] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10 (2009) R25.
[8] A.C. Lessinger, A.C. Martins Junqueira, T.A. Lemos, et al., The mitochondrial genome of the primary screwworm fly *Cochliomyia hominivorax* (Diptera: Calliphoridae). Insect Mol. Biol. 9 (2000) 521–529.
[9] B. Li, N. Fillmore, Y. Bai, et al., Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome Biol. 15 (2014) 553.
[10] M. Morgan, S. Anders, M. Lawrence, P. Aboyoun, H. Pagès, R. Gentleman, ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. Bioinformatics 25 (2009) 2607–2608.
[11] A. Oshlack, M.J. Wakefield, Transcript length bias in RNA-seq data confounds systems biology. Biol. Direct 4 (2009) 14.
[12] G. Parra, K. Bradnam, I. Korf, CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23 (2007) 1061–1067.
[13] M.D. Robinson, D.J. McCarthy, G.K. Smyth, edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26 (2010) 139–140.
[14] T.T. Torres, M. Dolezal, C. Schlötterer, B. Ottenwälder, Expression profiling of *Drosophila* mitochondrial genes via deep mRNA sequencing. Nucleic Acids Res. 37 (2009) 7509–7518.
[15] T.T. Torres, M. Metta, B. Ottenwalder, C. Schlötterer, Gene expression profiling by massively parallel sequencing. Genome Res. 18 (2008) 172–177.